Journal of Statistical Computation and Simulation
Vol. 00, No. 0, Month 2009, 1–17

Taylor & Francis
Taylor & Francis Group

Efficient estimators for adaptive two-stage sequential sampling

Mohammad Salehi[a]*, Mohammad Moradi[a], Jennifer A. Brown[b] and David Smith[c]

[a]*Department of Mathematical Sciences, Isfahan University of Technology, Isfahan, Iran;* [b]*Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand;* [c]*Leetown Science Center, US Geological Survey, 11649 Leetown Road, Kearneysville, WV 25430, USA* **Q1**

(*Received 16 January 2009; final version received 29 April 2009*)

In stratified sampling, methods for the allocation of effort among strata usually rely on some measure of within-stratum variance. If we do not have enough information about these variances, adaptive allocation can be used. In adaptive allocation designs, surveys are conducted in two phases. Information from the first phase is used to allocate the remaining units among the strata in the second phase. Brown *et al.* [*Adaptive two-stage sequential sampling*, Popul. Ecol. 50 (2008), pp. 239–245] introduced an adaptive allocation sampling design – where the final sample size was random – and an unbiased estimator. Here, we derive an unbiased variance estimator for the design, and consider a related design where the final sample size is fixed. Having a fixed final sample size can make survey-planning easier. We introduce a biased Horvitz–Thompson type estimator and a biased sample mean type estimator for the sampling designs. We conduct two simulation studies on honey producer in Kurdistan and synthetic zirconium distribution in a region on the moon. Results show that the introduced estimators are more efficient than the available estimators for both variable and fixed sample size designs, and the conventional unbiased estimator of stratified simple random sampling design. In order to evaluate efficiencies of the introduced designs and their estimator furthermore, we first review some well-known adaptive allocation designs and compare their estimator with the introduced estimators. Simulation results show that the introduced estimators are more efficient than available estimators of these well-known adaptive allocation designs.

Keywords: stratified population; Neyman's allocation; Horvitz–Thompson type estimator; sample mean type estimator **Q2**

AMS Subject Classification: 62D05; 92-08

1. Introduction

In conventional stratified sampling, the population is partitioned into regions or strata, and a simple random sample is selected in each stratum, with the selections in one stratum being independent of selections in the others. To obtain the best estimate of the total population with a given total **Q3** sample size or total survey cost, or to achieve a desired precision with minimum cost, optimal allocation of sample size among the strata is necessary, which results in larger sample sizes in strata that are larger, more variable and less costly to sample [1,2].

*Corresponding author. Email: salehi-m@cc.iut.ac.ir

ISSN 0094-9655 print/ISSN 1563-5163 online

DOI: 10.1080/00949650903005664
http://www.informaworld.com

If prior knowledge of the strata variances is not available, it would be natural to carry out the sampling in two phases and compute sample variances from the first phase, which are then used to adaptively allocate the remaining sample size among strata. Alternatively, allocation of the remaining sample size could be based on the stratum–sample mean or on the number of large values in the first-phase sample rather than sample variances, since with many natural populations high means or large values are associated with high variances. The standard stratified sampling estimator gives an unbiased estimator of the total population with conventional stratified random sampling, but it is not in general unbiased with adaptive allocation designs.

Francis [3] introduced an adaptive allocation method in stratified sampling. Francis suggested to select 75% of the final sample size by SRS, and then allocate the remaining sample among strata. In this method, the variances of estimators are estimated from the first-phase sample. Then, one unit is added to the stratum with the largest decrease in the variance. On the basis of the new sample set, the variance is estimated, and the second unit is selected from the stratum with the largest decrease in the variance. This process is continued to allocate the remaining units.

Jolly and Hampton [4] proposed another adaptive allocation. Their method is similar to Francis's method. At first, approximately 75% of the final sample size are selected and sample variances are estimated for all strata, and then, by Neyman's method, the remaining units are allocated to strata. They applied their method to an acoustic survey of South African anchovy.

Salehi and Smith [5] introduced two-stage sequential sampling. In the special case that all primary units are selected, the sampling design may be considered as an adaptive allocation sample design. In the first phase, a simple random sample is selected from each primary sample unit (stratum). To conduct the second phase, a condition C is defined for which the remaining sample size is allocated based on the value of the variable of interest. They used Murthy's estimator, which is an unbiased estimator, for the population mean.

Brown *et al.* [6] introduced an adaptive allocation with variable sample size design. The first phase is conducted similar to the first phase in Salehi–Smith's design. A multiplier d is considered before sampling. In the second phase, if l_{h1} units from the first phase sample in stratum h satisfied the condition C, then additional $d \times l_{h1}$ units are sampled from the remaining units in stratum h. They used Murthy's estimator to estimate the population mean.

In Section 2, we describe Brown *et al.*'s adaptive allocation sampling design and introduce a fixed sample size version. We also describe Francis's design, Jolly–Hampton's design, and Salehi–Smith's design. In Section 3, we derive an unbiased variance estimator for Brown *et al.* [6]. Two biased estimators are introduced for the variable and fixed sample size designs. We then study their sampling properties analytically. In Section 4, two empirical studies using honey producer population in Kurdistan and synthetic zirconium distribution population in a region on the moon, are described.

2. Sampling designs

Suppose that we have a population of N units which are partitioned into H strata of size N_h units ($h = 1, \ldots, H$). Let unit hi denote the ith unit in the hth stratum with an associated measurement or count y_{hi}, and K_h is the number of units satisfying a condition C in the hth stratum. The condition C for unit hi is defined as $(y_{hi} > c)$.

Suppose that variances of strata are unknown and we have no prior information on variances' estimator of strata. Therefore, Neyman's allocation cannot be used. We describe two adaptive allocation sampling designs, one with a variable sample size [6] and the other with a fixed sample size. We also describe adaptive allocation designs introduced by Francis [3], Jolly and Hompton [4], and Salehi and Smith [5] in this section.

2.1. *Variable sample size design*

In the variable sample size design, initially a first phase sample of size n_{h1} units is taken without replacement from each stratum where $n_1 = \sum_{h=1}^{H} n_{h1}$ is the total sample size of the first phase. In this paper we assume that we have no prior information about the strata variances, then in the variable sample size design and next designs we allocate a fixed n_{h1} to all strata proportional to stratum size. If l_{h1} units of the first phase sample units in the hth stratum satisfy the condition C, $d \times l_{h1}$ additional sample units are selected without replacement from the remaining units in stratum h. The multiplier d is chosen prior to sampling. Because sampling is without replacement in either phases, and d is a fixed multiplier in all strata, the following inequality can be imposed:

$$0 < d \leq \min \left\{ \frac{N_h - n_{h1}}{l_{h1}} | h = 1, \ldots, H \right\}.$$

When $n_{h1} + d \times l_{h1}$ is larger than N_h, we select all units in stratum h and we can, therefore, ignore the above restriction similar to Neyman's allocation.

In this design $n_2 = d \sum_{h=1}^{H} l_{h1}$, the number of adaptively added units, and, therefore, the final total sample size, $n = n_1 + n_2$, are random. This implies that

$$d = \frac{n - n_1}{\sum_h l_{h1}},$$

where n in this design is random. However, by fixing n prior to sampling, this relationship can be used to determine d under a fixed sample size version of the design.

2.2. *Fixed sample size design*

A variable sample size can make it difficult to plan and implement sampling. By setting the sample size prior to sampling and using the results from the first phase of sampling, a fixed sample size design can be introduced. Suppose that we fix the final sample size at n, and select a random sample of size n_{h1} without replacement from each of the strata. Then d will be bounded and is given by

$$d = \frac{n - n_1}{\sum_h l_{h1}}.$$

We can select $d \times l_{h1}$ units from stratum h when $\sum_h l_{h1} > 0$. In the fixed sample size design, if $\sum_h l_{h1} > 0$ the final sample size would be equal to the predetermined sample size n, but if $\sum_h l_{h1} = 0$, multiplier d is undefined. To achieve fixed sample size n, we allocate the remaining $n - \sum_h n_{h1}$ units equally to strata when $\sum_h l_{h1} = 0$.

We should note that attaining the predetermined sample size n is not strictly true because $n_1 + \sum_h d \times l_{h1}$ may not be a whole number. For example, consider this case where we want $n = 20$ and we have a population of three equal-sized strata. We start off with $n_1 = 9$. The three strata are sampled in the first phase by selecting three units in each, and we find $l_{h1} = 2$ in the first strata, 2 in the second and 1 in the third. The multiplier would be calculated as $d = (20 - 9)/(2 + 2 + 1) = 11/5 = 2.2$. We then select in the second phase $2.2 \times 2 = 4.4$ which we round to 4 units in the first stratum, 4 in the second and 2 in the third. Our final sample size is $9 + 4 + 4 + 2 = 19$ and not 20. The final sample size will be approximately equal to the predetermined sample size and may vary by a small amount depending on the effect of converting the real numbers to integers.

2.3. *Francis's design*

Francis [3] in his fisheries research allocated fixed sample size n to the strata in two phases. In the second phase, sampling was carried out in a sequential fashion. Francis's design is based on variance estimate of total weight of fish (biomass) in the first phase in stratum h. When we estimate the stratum total parameter by conventional estimator, we used Francis's design with little modification, as shown in the following. In the first phase, units are selected similar to the variable design. In this method to allocate remaining $n - \sum_{h=1}^{H} n_{h1}$ units, we should follow the following steps for each unit. If an additional unit is added to stratum h, then, using the same estimate s_{h1}^2 from the first phase, the reduction in the estimated variance of the conventional estimator is

$$G_h = N_h^2 \left(\frac{1}{n_{h1}} - \frac{1}{n_{h1}+1} \right) s_{h1}^2 = \frac{N_h^2 s_{h1}^2}{n_{h1}(n_{h1}+1)}.$$

This formula is now used to determine phase-2 allocations sequentially. The first unit of the remaining units is allocated to the stratum for which G_h is the greatest. Suppose that this is stratum j. Then G_j is recalculated as $s_{j1}^2/(n_{j1}+1)(n_{j1}+2)$. The next unit is added to the stratum for which G_h is a maximum, and so on.

2.4. *Jolly–Hampton's design*

Jolly and Hampton's adaptive allocation sampling design can be formulated as a fixed sample size design. In Jolly–Hampton's design, n_{h1} units has been selected from each stratum as previous designs. Suppose that the final sample size is fixed at n. A first phase sample of size n_{h1} is selected without replacement from stratum h such that $n_{h1} < n/H$. Then the remaining $n - n_1$ units are allocated as follows. Variances of strata are then estimated from the first phase sample. The sample size in the hth stratum is computed by

$$n_h = n_{h1} + (n - n_1) \frac{N_h s_{h1}}{\sum_{h=1}^{H} N_h s_{h1}},$$

where s_{h1} is the standard deviation of the first phase sample in the stratum h.

If sample size in the hth stratum is larger than N_h, we select all units in stratum h.

2.5. *Salehi–Smith's design*

Salehi–Smith's two-stage sequential sampling is an adaptive allocation sampling design when all primary sampling units are selected in the first phase. It can be formulated as a variable sample size design. In the first phase, a simple random sample of n_{h1} units is drawn without replacement from stratum $h (h = 1, \ldots, H)$. If condition C is satisfied for at least one unit in the hth stratum in the first phase sample, a predetermined number of additional units, say n_{h2}, are selected at random from the remaining units in stratum h. As a result $n = \sum_h n_{h1} + \sum_h n_{h2} = n_1 + n_2$ is the sample size and is a random value.

2.6. *Comparison of variable and fixed sample size designs*

The variable sample size design and the Salehi–Smith's design share an advantage over the other two designs, in that the allocation of second phase effort can be done during the first phase. The fixed sample size design, the Francis design, and the Jolly–Hampton design all require the first phase to be completed and strata to be visited before second phase allocation can occur. This

means that stratum that is to be surveyed in the second phase will need to be revisited, and this may be costly.

Conversely, the variable sample size design and the Salehi–Smith's design share a disadvantage over the other designs in that the final sample size is not known prior to surveying. This can make planning the survey difficult. With the fixed sample size design, the Francis design, and the Jolly–Hampton design, the size of the final sample is known.

3. Estimations

Let $\tau_h = \sum_{i=1}^{N_h} y_{hi}$ be the total of y values in the hth stratum, and $\tau = \sum_{h=1}^{H} \tau_h$ be the total population. To estimate τ and $\mathrm{var}(\hat{\tau})$, we estimate τ_h and $var(\hat{\tau}_h)$. We then have

$$\hat{\tau} = \sum_{h=1}^{H} \hat{\tau}_h$$

$$\widehat{\mathrm{var}}(\hat{\tau}) = \sum_{h=1}^{H} \widehat{\mathrm{var}}(\hat{\tau}_h).$$

We will derive Murthy's estimator for the variable sample size design [6], and two biased estimators for both fixed and random sample size designs. The first biased estimator – the Horvitz–Thompson type estimator ($\hat{\tau}_{\mathrm{HT}}$) – is the HT estimator for which the inclusion probabilities are estimated. The HT estimator is an unbiased estimator but the inclusion probabilities depend on the parameters of population for the introduced sampling designs. We, therefore, estimate the inclusion probabilities, so that $\hat{\tau}_{\mathrm{HT}}$ is a biased estimator. The second biased estimator is the sample mean type estimator.

3.1. *Murthy's estimator in the variable sample size design*

Brown *et al.* [6] introduced Murthy's estimator for the variable sample size design, which is an unbiased estimator

$$\hat{\tau}_{\mathrm{M}h} = \sum_{i=1}^{n} \frac{P(\mathcal{S}_h|i)}{P(\mathcal{S}_h)} y_{hi} = N_h[\hat{p}_h \bar{y}_{ch} + (1 - \hat{p}_h)\bar{y}_{c'h}].$$

We derive its variance estimator, which is given by

$$\mathrm{var}(\hat{\tau}_{\mathrm{M}h}) = \sum_{i=1}^{N_h} \sum_{j<i}^{N_h} \left(1 - \sum_{\mathcal{S}_h \ni i,j} \frac{P(\mathcal{S}_h|i)P(\mathcal{S}_h|j)}{P(\mathcal{S}_h)}\right)(y_{hi} - y_{hj})^2,$$

where $P(\mathcal{S}_h)$ is the probability of getting sample $\mathcal{S}_h$, $P(\mathcal{S}_h|i)$ is the conditional probability of getting the sample $\mathcal{S}_h$, given the ith unit was selected first, $\bar{y}_{ch}$ and $\bar{y}_{c'h}$ are respectively the mean of units satisfying the condition C and not satisfying the condition C in stration h, and $\hat{p}_h = l_{h1}/n_{h1}$.

It can be shown that, using Rao–Blackwell method, this estimator is an improved estimator and it is a function of minimal sufficient statistics $\mathcal{S}_h$, where $\mathcal{S}_h$ is the observed unordered set of distinct units in the sample.

From the Rao–Blackwell theorem $\mathrm{var}(\hat{\tau}_{Mh}) = \mathrm{var}(\bar{y}_{h1}) - E(\mathrm{var}(\bar{y}_{h1})|\mathcal{S}_h)$ and $\widehat{\mathrm{var}}(\hat{\tau}_{\mathrm{M}h}) = \widehat{\mathrm{var}}(\bar{y}_{h1}) - E(\widehat{\mathrm{var}}(\bar{y}_{h1})|\mathcal{S}_h)$.

With routine but lengthy algebra, $\widehat{\mathrm{var}}[\hat{\tau}_{\mathrm{M}h}]$ is simplified as follows:

$$\begin{aligned}\widehat{\mathrm{var}}(\hat{\tau}_{\mathrm{M}h}) = N_h \Bigg\{ & \hat{p}_h \left[\frac{(N_h-1)(l_{h1}-1)}{n_{h1}-1} + \left(\frac{(N_h-n_{h1})(1-\hat{p}_h)}{n_{h1}-1} - N_h \hat{p}_h \right) \frac{l_h-1}{l_h} \right] s_{ch}^2 \\ & + \hat{p}_h(1-\hat{p}_h) \frac{N_h-n_{h1}}{n_{h1}-1} (\bar{y}_{ch} - \bar{y}_{c'h})^2 + (1-\hat{p}_h) \\ & \times \left[\frac{(N_h-1)(n_{h1}-l_{h1}-1)}{n_{h1}-1} + \left(\frac{(N_h-n_{h1})\hat{p}_h}{n_{h1}-1} - N_h(1-\hat{p}_h) \right) \frac{l'_h-1}{l'_h} \right] s_{c'h}^2 \Bigg\},\end{aligned}$$

where $\mathcal{S}_{ch}$ and $\mathcal{S}_{c'h}$ are, respectively, the subsamples of units satisfying the condition C and the subsamples of units not satisfying the condition C in stratum h, l_h and l'_h are, respectively, the cardinality of these subsamples in stratum h, $s_{ch}^2 = 1/(l_h-1)\sum_{i\in\mathcal{S}_{ch}}(y_{hi}-\bar{y}_{ch})^2$, $s_{c'h}^2 = 1/(l'_h-1)\sum_{i\in\mathcal{S}_{c'h}}(y_{hi}-\bar{y}_{c'h})^2$.

Brown *et al.* [6] showed that this estimator is efficient than the estimator for a very rare population. We should note that using Rao–Blackwell method would improve the estimator more when the second phase sample is more consistent with the first phase sample.

3.2. *Horvitz–Thompson type estimator*

3.2.1. *Horvitz–Thompson type estimator in variable sample size design*

Horvitz and Thompson [7] introduced an unbiased estimator for the total population, which we can apply it for estimating τ_h as

$$\hat{\tau}_{\mathrm{HT}h} = \sum_{i\in\mathcal{S}_h} \frac{y_{hi}}{\pi_{hi}},$$

where π_{hi} is the inclusion probability of the hith unit in the sample. In the introduced variable sample size design, π_{hi} is given by

$$\pi_{hi} = \begin{cases} \displaystyle\sum_{l_{h1}=1}^{n_{h1}} \frac{\binom{K_h-1}{l_{h1}-1}\binom{N_h-K_h}{n_{h1}-l_{h1}}\binom{N_h-n_{h1}}{d\times l_{h1}} + \binom{K_h-1}{l_{h1}}\binom{N_h-K_h}{n_{h1}-l_{h1}}\binom{N_h-n_{h1}-1}{d\times l_{h1}-1}}{\binom{N_h}{n_{h1}}\binom{N_h-n_{h1}}{d\times l_{h1}}}, & i\in\mathcal{S}_{ch}, \\ \displaystyle\sum_{l_{h1}=0}^{n_{h1}-1} \frac{\binom{K_h}{l_{h1}}\binom{N_h-K_h-1}{n_{h1}-l_{h1}-1}\binom{N_h-n_{h1}}{d\times l_{h1}}}{\binom{N_h}{n_{h1}}\binom{N_h-n_{h1}}{d\times l_{h1}}} \\ \quad + \displaystyle\sum_{l_{h1}=1}^{n_{h1}} \frac{\binom{K_h}{l_{h1}}\binom{N_h-K_h-1}{n_{h1}-l_{h1}}\binom{N_h-n_{h1}-1}{d\times l_{h1}-1}}{\binom{N_h}{n_{h1}}\binom{N_h-n_{h1}}{d\times l_{h1}}}, & i\in\mathcal{S}_{c'h}, \end{cases}$$

which can be simplified as

$$\pi_{hi} = \begin{cases} \dfrac{n_{h1}}{N_h}\left(1 + d\left(\dfrac{K_h - 1}{N_h - 1}\right)\right) & \text{if } i \in \mathcal{S}_{ch}, \\ \dfrac{n_{h1}}{N_h}\left(1 + d\left(\dfrac{K_h}{N_h - 1}\right)\right) & \text{if } i \in \mathcal{S}_{c'h}, \end{cases}$$

where K_h is the number of units satisfying the condition C in stratum h. It cannot be calculated, given $\mathcal{S}$. Inclusion probability π_{hi} depends on parameter K_h. Christman [8] introduced adaptive two-stage one-per-stratum sampling in which she used the same HT estimator for estimating its component. Her empirical study showed that the introduced estimator is efficient.

We use unbiased and consistent estimator $\hat{K}_h = N_h l_{h1}/n_{h1}$ in π_{hi}. We then have

$$\hat{\tau}_h = \sum_{i \in \mathcal{S}_h} \frac{y_{hi}}{\hat{\pi}_{hi}},$$

where

$$\hat{\pi}_{hi} = \begin{cases} \dfrac{n_{h1}}{N_h}\left(1 + d\left(\dfrac{\hat{K}_h - 1}{N_h - 1}\right)\right) & \text{if } i \in \mathcal{S}_{ch}, \\ \dfrac{n_{h1}}{N_h}\left(1 + d\left(\dfrac{\hat{K}_h}{N_h - 1}\right)\right) & \text{if } i \in \mathcal{S}_{c'h}. \end{cases}$$

The variance estimator is given by

$$\widehat{\text{var}}(\hat{\tau}_{\text{HT}h}) = \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{S}} \left(\frac{\pi_{ij}}{\pi_i \pi_j} - 1\right) \frac{y_i y_j}{\pi_{ij}},$$

where π_{ij} is the joint inclusion probability which is given by

$$\pi_{ij} = \begin{cases} \dfrac{n_{h1}(n_{h1}-1)}{N_h(N_h-1)}\left\{1 + \dfrac{d}{n_{h1}-1}\left(2 + \dfrac{(2n_{h1}-3)(K_h-2)}{N_h-2} + \dfrac{d(K_h-2)}{(N_h-2)^2}\left[\dfrac{(N_h-K_h)(N_h-n_{h1}-2)}{N_h-3} + n_{h1}(K_h-2)\right]\right)\right\} & \text{if } i, j \in \mathcal{S}_{ch}, \\ \dfrac{n_{h1}(n_{h1}-1)}{N_h(N_h-1)}\left\{1 + \dfrac{d}{n_{h1}-1}\left(1 + \dfrac{(2n_{h1}-3)(K_h-1)}{N_h} + \dfrac{d(K_h-1)}{(N_h-2)^2}\left[\dfrac{(N_h-K_h-1)(N_h-n_{h1}-2)}{N_h-3} + n_{h1}(K_h-1)\right]\right)\right\} & \text{if } i \in \mathcal{S}_{ch}, \; j \in \mathcal{S}_{c'h}, \\ \dfrac{n_{h1}(n_{h1}-1)}{N_h(N_h-1)}\left\{1 + \dfrac{d}{n_{h1}-1}\left(\dfrac{(2n_{h1}-3)K_h}{N_h} + \dfrac{dK_h}{(N_h-2)^2}\left[\dfrac{(N_h-K_h-2)(N_h-n_{h1}-2)}{N_h-3} + n_{h1}K_h\right]\right)\right\} & \text{if } i, j \in \mathcal{S}_{c'h}. \end{cases}$$

Once again π_{ij} is a function of population parameter K_h, so we use $\hat{\pi}_{ij}$.

3.2.2. *Horvitz–Thompson type estimator in fixed sample size design*

Inclusion probability in fixed sample size design is more complicated because multiplier d is not fixed before completing the first phase. It depends on the number of sample units satisfying the condition C in the first phase. We now estimate the inclusion probability in fixed sample size design with respect to variable D.

For fixed sample size, design D is as follows:

$$D = \begin{cases} \dfrac{n-n_1}{\sum_h l_{h1}} & \text{if } \displaystyle\sum_h l_{h1} \in \{1,2,\ldots,n_1\}, \\ \dfrac{n-n_1}{H} & \text{if } \displaystyle\sum_h l_{h1} = 0. \end{cases}$$

$$\pi'_{hi} = \sum_{\mathcal{S}\ni i} P(\mathcal{S}) = \sum_{\mathcal{S}\ni i} \sum_d P(\mathcal{S}, D=d) = \sum_d P(D=d) \sum_{\mathcal{S}\ni i} P(\mathcal{S}|D=d).$$

We know from the previous design that the second summation in the above equation is the inclusion probability in the variable sample size design, so that

$$\pi'_{hi} = \sum_d P(D=d)\pi_{hi} = \begin{cases} \displaystyle\sum_d P(D=d)\frac{n_{h1}}{N_h}\left(1+d\frac{K_h-1}{N_h-1}\right) & i \in \mathcal{S}_{ch}, \\ \displaystyle\sum_d P(D=d)\frac{n_{h1}}{N_h}\left(1+d\frac{K_h}{N_h-1}\right) & i \in \mathcal{S}_{c'h}. \end{cases}$$

$$= \begin{cases} \dfrac{n_{h1}}{N_h}\left(1+E(D)\dfrac{K_h-1}{N_h-1}\right) & i \in \mathcal{S}_{ch}, \\ \dfrac{n_{h1}}{N_h}\left(1+E(D)\dfrac{K_h}{N_h-1}\right) & i \in \mathcal{S}_{c'h}. \end{cases}$$

To calculate $E(D)$ we need to have probability function of D, where it is

$$P(D=d) = \begin{cases} P\left(\sum_h l_{h1} = \dfrac{n-n_1}{d}\right) & \text{if } d \neq \dfrac{n-n_1}{H}, \\ P(\sum_h l_{h1} = 0) & \text{if } d = \dfrac{n-n_1}{H}. \end{cases}$$

$$P\left(\sum_h l_{h1} = l.\right) = \sum_{l_1,\ldots,l_{H-1}} P(l_1, l_2, \ldots, l_{H-1}, l. - l_1 - l_2 - \cdots - l_{H-1})$$

$$= \sum_{l_1,\ldots,l_{H-1}} P(l_1)P(l_2)\cdots P(l_{H-1})P(l. - l_1 - l_2 - \cdots - l_{H-1})$$

$$= \sum_{l_1,\ldots,l_{H-1}} \frac{\binom{k_1}{l_1}\binom{N_1-K_1}{n_{h1}-l_1}}{\binom{N_1}{n_{h1}}} \cdots$$

$$\times \frac{\binom{K_H}{l. - l_1 - \cdots - l_{H-1}}\binom{N_H - K_H}{n_{hH} - (l. - l_1 - \cdots - l_{H-1})}}{\binom{N_H}{n_{hH}}}$$

The probability function of D depends on the population parameters, and calculating $E(D)$ from the sample is impossible. To solve the problem we recommend to replace the unbiased estimator D with $E(D)$. With this substitution, equation of inclusion probabilities in both variable and fixed sample size design will be identical.

3.3. *Sample mean type estimator of total*

When N_h is large, we have

$$\lim_{N_h\to\infty} \frac{dn_{h1}}{N_h(N_h-1)} = 0.$$

Thus

$$\begin{aligned}\lim_{N_h\to\infty}(\hat{\tau}_h) &= \lim_{N_h\to\infty}\left(\frac{y_{.ch}}{n_{h1}/N_h + dl_{h1}/(N_h-1) - dn_{h1}/N_h(N_h-1)} + \frac{y_{.c'h}}{n_{h1}/N_h + dl_{h1}/(N_h-1)}\right)\\ &\cong \frac{y_{.ch}+y_{.c'h}}{(n_{h1}+dl_h)/N_h} = N_h\bar{y}_{hs},\end{aligned}$$

where $y_{.ch}$ and $y_{.c'h}$ are, respectively, the total of units satisfying the condition C and not satisfying C. We, therefore, can use estimator $\hat{\tau}_s = \sum_h N_h\bar{y}_{hs}$, when N_hs are relative large. It is clear that $\hat{\tau}_s$ is another biased estimator which can be computed easily. When N_h is large, the sample mean and HT type estimator are approximately equal but when N_h is small they can be different estimators. In the appendix, we show that $\hat{\tau}_s$ is a negatively biased estimator.

4. Simulation study

In this section, the efficiency of the five sample designs are investigated using two populations. The first population is honey producer in Kurdistan Province of Iran (HPK) [9], and the second simulated population synthetic zirconium in a region on the Moon (SZM) [10]. Since we did not have access to the value of units not satisfying condition C we simulate them. The results for both populations are almost similar but we present both to illustrate the versatility of the sampling designs. We also want to use the most recently used populations rather than the most frequent populations in this context. The maps of HPK and SZM are shown in Figures 1 and 2, respectively. In population HPK, cities or villages are sampling units of which there were 1864. We partitioned the map of Kurdistan province into four strata, each containing 466 units. The units in each strata have approximately the same geographical condition. We are interested in the variable y_i, the amount of produced honey in unit i in a year. We defined the condition C ($y_i > 20$) which can be considered as active units in producing honey. In HPK, K_h is 0, 1, 15, and 39 in the four strata. For the SZM population [10], the region of the moon is partitioned into 560 equal-sized units and synthetic zirconium is measured in each. The region is stratified into eight strata. The condition is defined as $y_i \geq 1$. In SZM, K_h is 3, 0, 11, 0, 22, 0, 9, and 0 in the eight strata.

We use the five sample designs introduced in Section 2 and compare the efficiency of designs with different estimators (Section 3) by Monte Carlo simulation. We use 30,000 replicate samples for each design and estimator combination. Efficiency was estimated by comparing the design and estimator combination with stratified simple random sampling using the same sample size within each stratum. We estimate the relative bias of introduced biased estimators in the five introduced sample designs by the Monte Carlo simulation.

For the variable sample size design, the set of first phase sample sizes in each stratum was $n_{h1} = (2, 3, \ldots, 10, 20, \ldots, 50, 100, 150, 200)$ and $d = (1, 2, 3, 4)$. For the fixed sample size

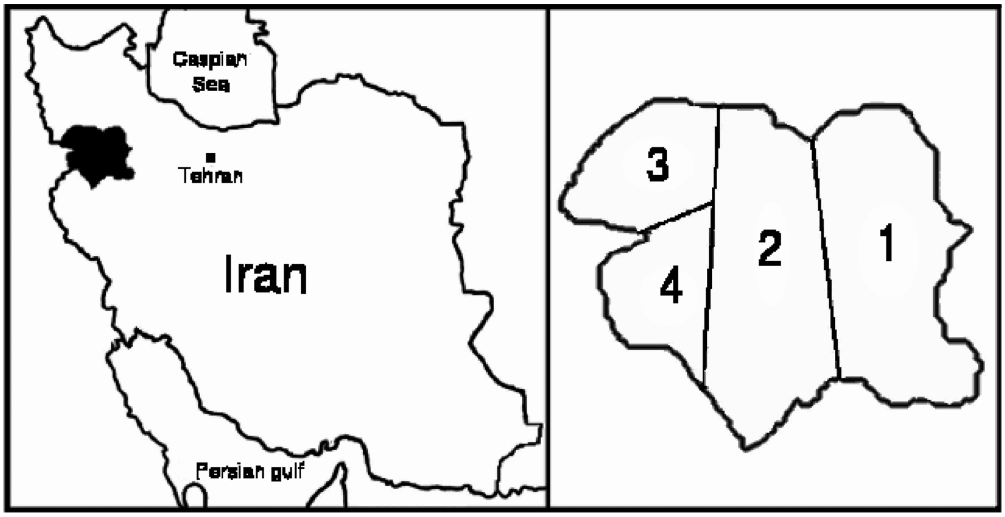

Figure 1. The shaded in region black on the map is Kurdistan province of Iran that is partitioned to four strata and numbers show the strata.

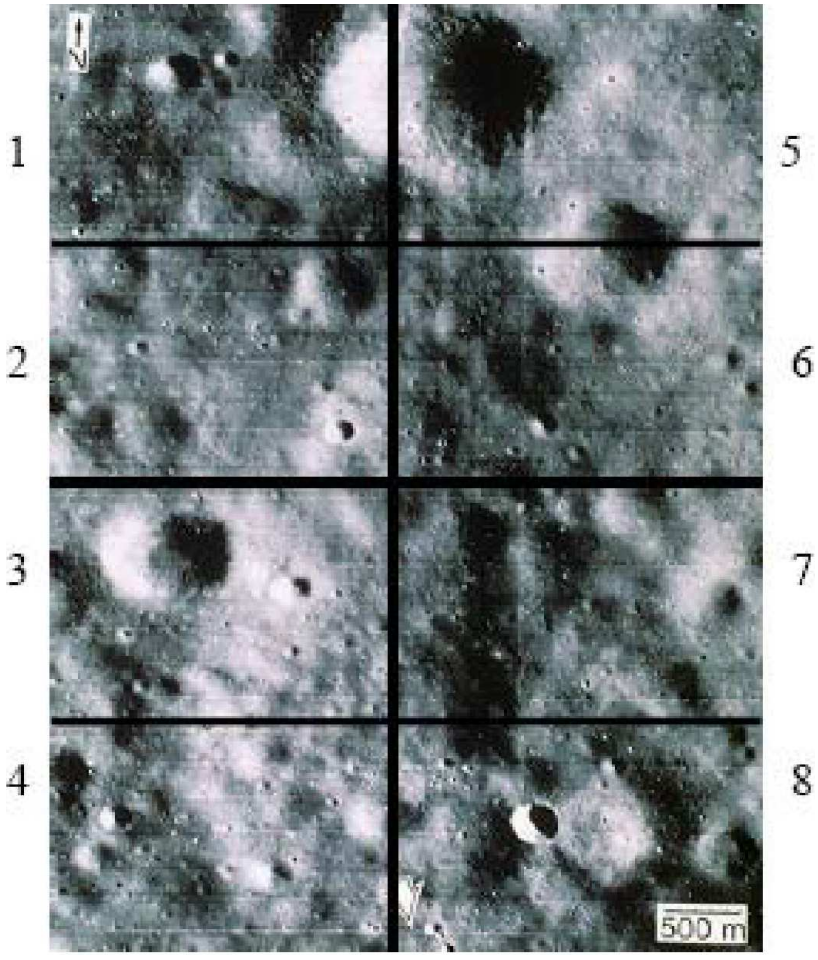

Figure 2. Lunar Orbiter photograph III-133-H2 of prospecting region north of Apollo 14 landing site (photograph courtesy of Lunar and Planetary Institute). Bold numbers in the two besides are the strata numbers.

design, the Francis design, and for the Jolly–Hampton's design, the final sample size was selected so that it equalled the effective sample size for all pairs of n_{h1} and d from the variable sample size design. For the Salehi–Smith's design, we selected n_2 such that the effective sample size was matched with the effective sample size for all pairs of n_{h1} and d from the variable sample size design.

The sample mean type estimator was used for all five designs. In addition, the HT type estimator was used for the variable and fixed sample size design, and Murthy's estimator for the variable sample size design, for the fixed sample size design and for the Salehi–Smith's design (see [5] for details of Murthy's estimator).

For each combination of designs and estimator we calculated

$$MSE[\hat{\tau}_\star] = \frac{1}{29999}\sum_{i=1}^{30000}(\hat{\tau}_\star - \bar{\tau}_\star)^2 + (\bar{\tau}_\star - \tau)^2,$$

where $\bar{\tau}_\star = 1/30000\sum_{i=1}^{30000}\hat{\tau}_{i\star}$. We have used the notation $\hat{\tau}_\star$, which stands for $\hat{\tau}_{\mathrm{M}}$, $\hat{\tau}_{\mathrm{HT}}$, $\hat{\tau}_s$, where $\hat{\tau}_{\mathrm{M}} = \sum_h \hat{\tau}_{\mathrm{M}h}$, $\hat{\tau}_{\mathrm{HT}} = \sum_h \hat{\tau}_h$, and $\hat{\tau}_s = \sum_h N_h\bar{y}_{hs}$.

The relative efficiency of $\hat{\tau}_\star$ is given by

$$\mathrm{eff}[\hat{\tau}_\star] = \frac{\mathrm{VAR}[\hat{\tau}_{st}]}{\mathrm{MSE}[\hat{\tau}_\star]} \times 100.$$

For each combination of designs and biased estimators for HPK we calculated

$$RB(\hat{\tau}_\star) = \frac{\bar{\tau}_\star - \tau}{\tau}$$

where $\hat{\tau}_\star$ stands for $\hat{\tau}_{\mathrm{HT}}$ and $\hat{\tau}_s$.

Results are summarized in Tables 1–3, and in Figures 3–8, where e.s.s, v.d, f.d, Fr.d, J.d and S.d are used as abbreviations of effective sample size design, variable sample size design, fixed sample size design, Francis's design, Jolly–Hampton's design and Salehi–Smith's design.

In Figures 3– 8, the efficiency of estimators for the variable designs (v.d and S.d), fixed designs (f.d, Fr.d, and J.d), and the best estimators in the variable and fixed designs for two populations are plotted. From Tables 1 and 2, we can see the efficiency of the sample mean type and HT type for

Table 1. Simulation of efficiencies in the honey producer population in Kurdistan province.

n_h	d	e.s.s	$\hat{\tau}_{\mathrm{HT}}$-v.d	$\hat{\tau}_s$-v.d	$\hat{\tau}_{\mathrm{HT}}$-f.d	$\hat{\tau}_s$-f.d	$\hat{\tau}_{\mathrm{M}}$-v.d	$\hat{\tau}_s$-Fr.d	$\hat{\tau}_s$-J.d	$\hat{\tau}_{\mathrm{M}}$-S.d	$\hat{\tau}_s$-S.d
5	1	20.6	116	116	103	106	98	106	100	102	111
10	1	41.1	110	110	102	102	98	103	104	103	106
50	1	205.6	106	106	104	104	100	103	102	103	101
100	1	411	104	104	103	103	99	104	102	101	102
200	1	822	104	104	105	105	99	105	103	102	102
5	2	21.1	126	126	107	108	96	106	108	104	122
10	2	42.2	115	115	106	104	98	107	102	103	108
50	2	211.1	107	108	105	105	99	108	103	102	103
100	2	422.3	115	115	106	104	98	107	102	103	104
200	2	844.6	109	109	111	110	98	112	107	104	104
5	3	21.7	129	130	108	110	93	111	109	101	129
10	3	43.34	119	119	105	105	95	108	106	103	113
50	3	216.8	110	111	108	108	99	108	106	106	105
100	3	433.5	109	110	111	111	100	109	102	106	106
200	3	866.9	115	115	111	111	99	115	106	108	108
5	4	22.2	131	131	113	110	93	112	113	100	132
10	4	44.5	120	120	113	112	100	113	110	103	115
50	4	222.3	112	112	111	111	99	110	102	106	105
100	4	444.5	113	114	113	113	98	113	107	110	110
200	4	889.2	117	118	119	118	96	120	110	110	110

The population is partitioned into four strata and the condition C is $y_{hi} > 20$. The estimators $\hat{\tau}_{\mathrm{HT}}$, $\hat{\tau}_s$, and $\hat{\tau}_{\mathrm{M}}$ are, respectively, the HT type estimator, sample mean type estimator and Murthy's estimator. The notations e.s.s, v.d, f.d, Fr.d, J.d and S.d are, respectively, the effective sample size, variable sample size design, fixed sample size design, Francis's design, Jolly–Hompton's design and Salehi–Smith's design.

Table 2. Simulation of efficiencies in the synthetic zirconium distribution population in a region on the moon.

n_h	d	e.s.s	$\hat{\tau}_{\mathrm{HT}}$-v.d	$\hat{\tau}_s$-v.d	$\hat{\tau}_{\mathrm{HT}}$-f.d	$\hat{\tau}_s$-f.d	$\hat{\tau}_{\mathrm{M}}$-v.d	$\hat{\tau}_s$-Fr.d	$\hat{\tau}_s$-J.d	$\hat{\tau}_{\mathrm{M}}$-S.d	$\hat{\tau}_s$-S.d
5	1	42.87	115	113	106	105	96	105	106	103	107
10	1	85.7	111	110	105	105	95	105	102	102	101
20	1	171.5	110	112	108	108	95	108	103	109	108
30	1	257.1	110	113	110	110	94	113	104	110	109
5	2	45.7	115	118	107	106	91	107	106	105	110
10	2	91.4	113	113	112	111	93	104	104	106	105
20	2	182.8	114	115	116	115	91	111	106	111	110
30	2	274.2	122	121	123	122	90	118	108	118	117
5	3	48.5	113	114	107	108	91	105	104	106	111
10	3	97.1	110	113	111	110	88	102	106	110	109
20	3	194.3	119	115	121	119	86	109	108	115	114
30	3	291.5	128	126	131	130	83	124	111	123	123
5	4	51.5	109	108	105	104	85	103	109	100	106
10	4	103	110	112	110	109	84	103	104	111	109
20	4	206	120	118	122	119	82	111	111	121	118
30	4	308.1	135	135	140	137	78	130	113	134	133

The population is partitioned into eight strata and the condition C is $y_{hi} \geq 1$. The estimators $\hat{\tau}_{\mathrm{HT}}$, $\hat{\tau}_s$, and $\hat{\tau}_{\mathrm{M}}$ are, respectively, the HT type estimator, sample mean type estimator and Murthy's estimator. The notations e.s.s, v.d, f.d, Fr.d, J.d and S.d are, respectively, the effective sample size, variable sample size design, fixed sample size design, Francis's design, Jolly–Hompton's design and Salehi–Smith's design.

Table 3. Simulation of relative biases in HPK.

n_h	d	e.s.s	$\hat{\tau}_{\mathrm{HT}}$-v.d	$\hat{\tau}_s$-v.d	$\hat{\tau}_{\mathrm{HT}}$-f.d	$\hat{\tau}_s$-f.d	$\hat{\tau}_s$-Fr.d	$\hat{\tau}_s$-J.d	$\hat{\tau}_s$-S.d
5	1	20.6	−0.02	−0.02	−0.009	−0.015	−0.013	−0.015	−0.022
10	1	41.1	−0.01	−0.01	−0.008	−0.009	−0.006	−0.004	−0.008
50	1	205.6	−0.001	−0.003	−0.001	−0.001	−0.002	−0.0007	−0.0012
100	1	411	−0.0005	−0.0008	−0.0009	−0.001	−0.0004	−0.0003	−0.0005
200	1	822	−0.00004	−0.00007	−0.0003	−0.00002	−0.0002	−0.0002	−0.0005
5	2	21.1	−0.04	−0.04	−0.017	−0.017	−0.013	−0.018	−0.032
10	2	42.2	−0.02	−0.02	−0.014	−0.013	−0.008	−0.005	−0.018
50	2	211.1	−0.01	−0.01	−0.002	−0.002	−0.005	−0.0016	−0.0017
100	2	422.3	−0.001	−0.002	−0.001	−0.001	−0.001	−0.0008	−0.0003
200	2	844.6	−0.001	−0.001	−0.0003	−0.0009	−0.00002	−0.00003	−0.0002
5	3	21.7	−0.04	−0.04	−0.025	−0.03	−0.029	−0.02	−0.047
10	3	43.34	−0.03	−0.03	−0.017	−0.016	−0.014	−0.008	−0.022
50	3	216.8	−0.01	−0.01	−0.002	−0.003	−0.004	−0.0017	−0.002
100	3	433.5	−0.0006	−0.003	−0.001	−0.002	−0.002	−0.0003	−0.0005
200	3	866.9	−0.001	−0.001	−0.0003	−0.0005	−0.0005	−0.00006	−0.0002
5	4	22.2	−0.06	−0.06	−0.032	−0.029	−0.028	−0.019	−0.052
10	4	44.5	−0.03	−0.03	−0.018	−0.016	−0.02	−0.009	−0.029
50	4	222.3	−0.01	−0.01	−0.002	−0.003	−0.003	−0.003	−0.0022
100	4	444.5	−0.003	−0.004	−0.001	−0.002	−0.002	−0.002	−0.0007
200	4	889.2	−0.001	−0.002	−0.0005	−0.0005	−0.001	−0.00003	−0.0002

The population is partitioned into four strata and the condition C is $y_{hi} > 20$. The estimators $\hat{\tau}_{\mathrm{HT}}$ and $\hat{\tau}_s$ are, respectively, the HT type estimator and sample mean type estimator. The notations e.s.s, v.d, f.d, Fr.d, J.d and S.d are, respectively, the effective sample size, variable sample size design, fixed sample size design, Francis's design, Jolly–Hampton's design and Salehi–Smith's design.

the variable and fixed designs are approximately identical. To make clearer the plots in Figures 3–8 we only draw the sample mean type for the variable sample size design and fixed sample size design. Figures 3 and 6 show that the sample mean type for the variable sample size design is more efficient than other estimators in variable designs. For Salehi–Smith's design, when n_{h1} is small (for HPK $n_{h1} < 20$ and for SZM $n_{h1} < 10$) the sample mean type is more efficient than Murthy's estimator, but when n_{h1} is large, the estimators are approximately identical. In Figure 4, we can see the sample mean type for the fixed sample size design and Francis's design has approximately

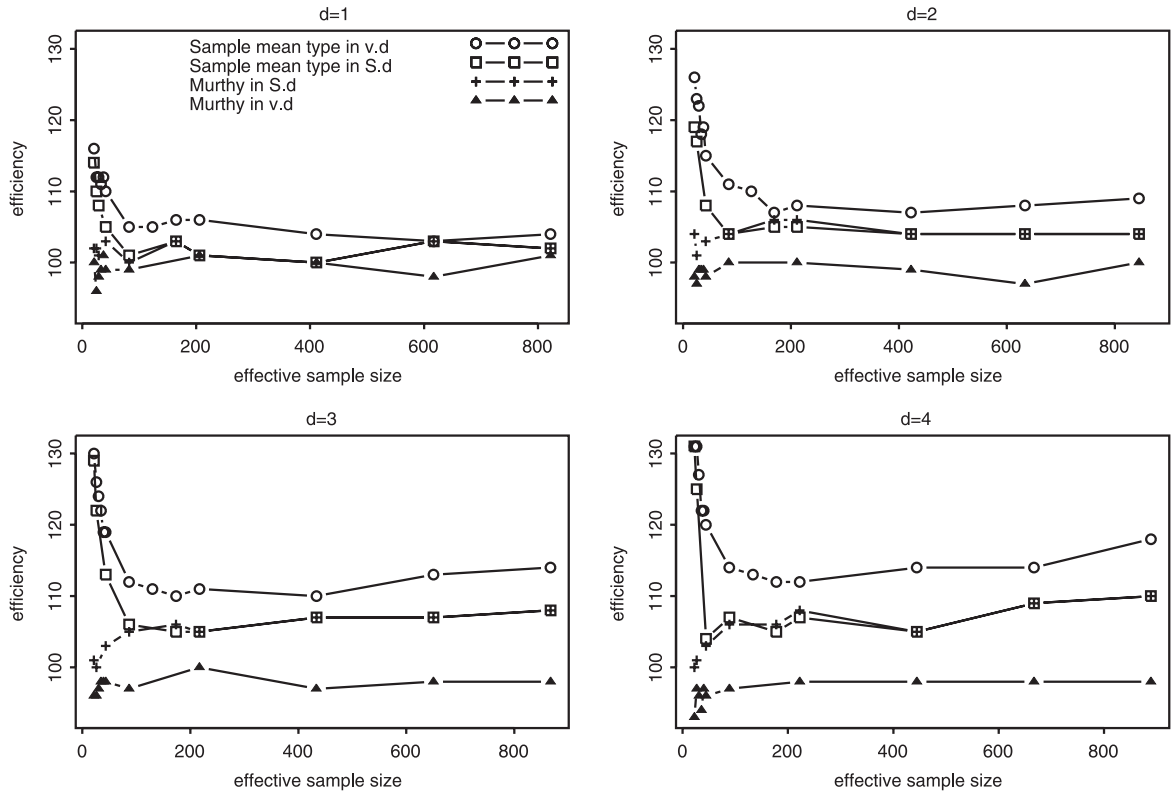

Figure 3. Efficiency of the sample mean type and Murthy's estimators in the variable sample size designs in the HPK population. The first phase sample size in each stratum is represented respectively as the set $\{5, 6, \ldots, 10, 20, \ldots, 50, 100, 200\}$.

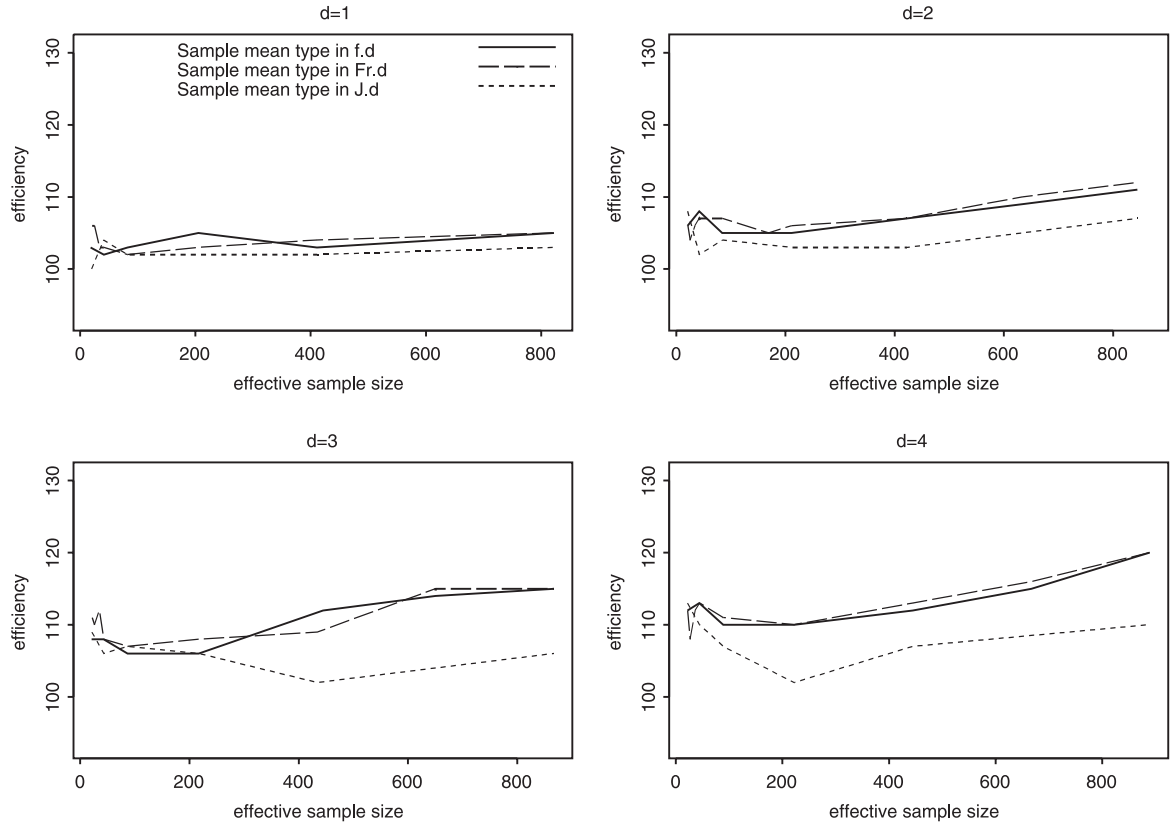

Figure 4. Efficiency of the sample mean type estimator in the fixed sample size designs in the HPK population. The first phase sample size in each stratum is represented respectively as the set $\{5, 6, \ldots, 10, 20, \ldots, 50, 100, 200\}$.

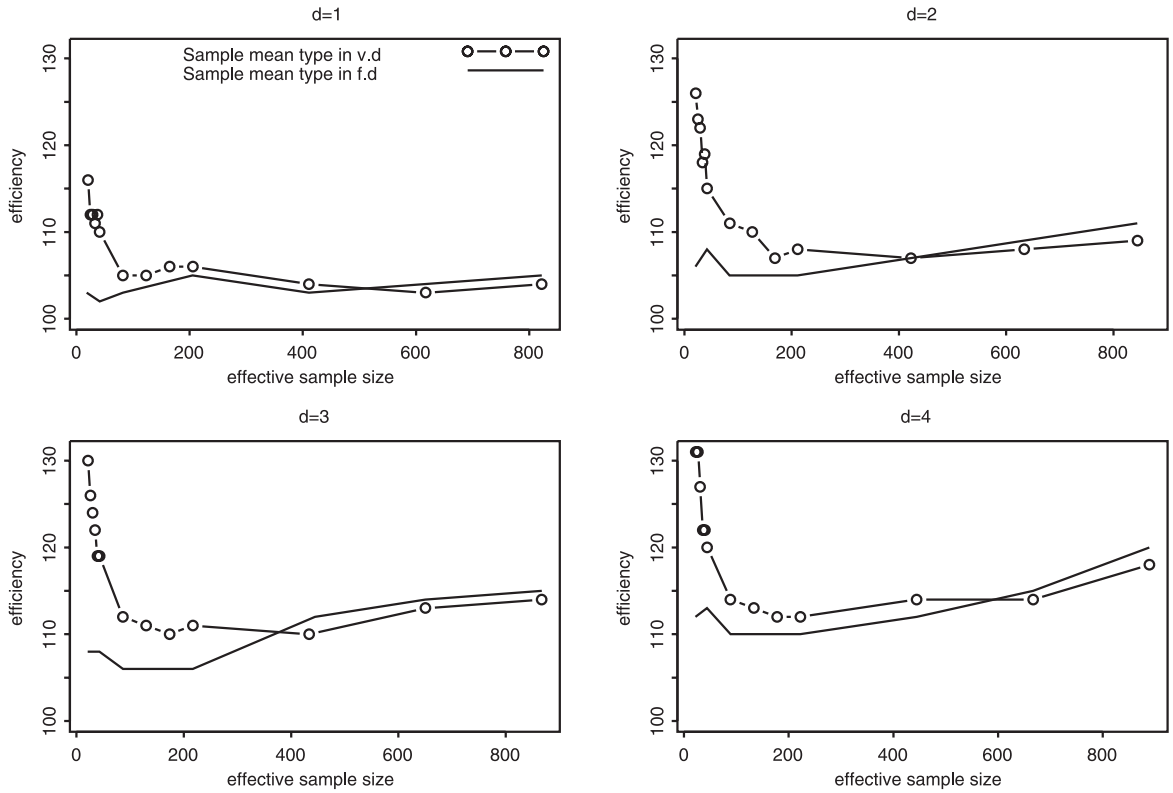

Figure 5. Efficiency of the best estimators in all designs in the HPK population. The first phase sample size in each stratum is represented respectively as the set {5, 7 . . . , 10, 20, . . . , 50, 100, 200}.

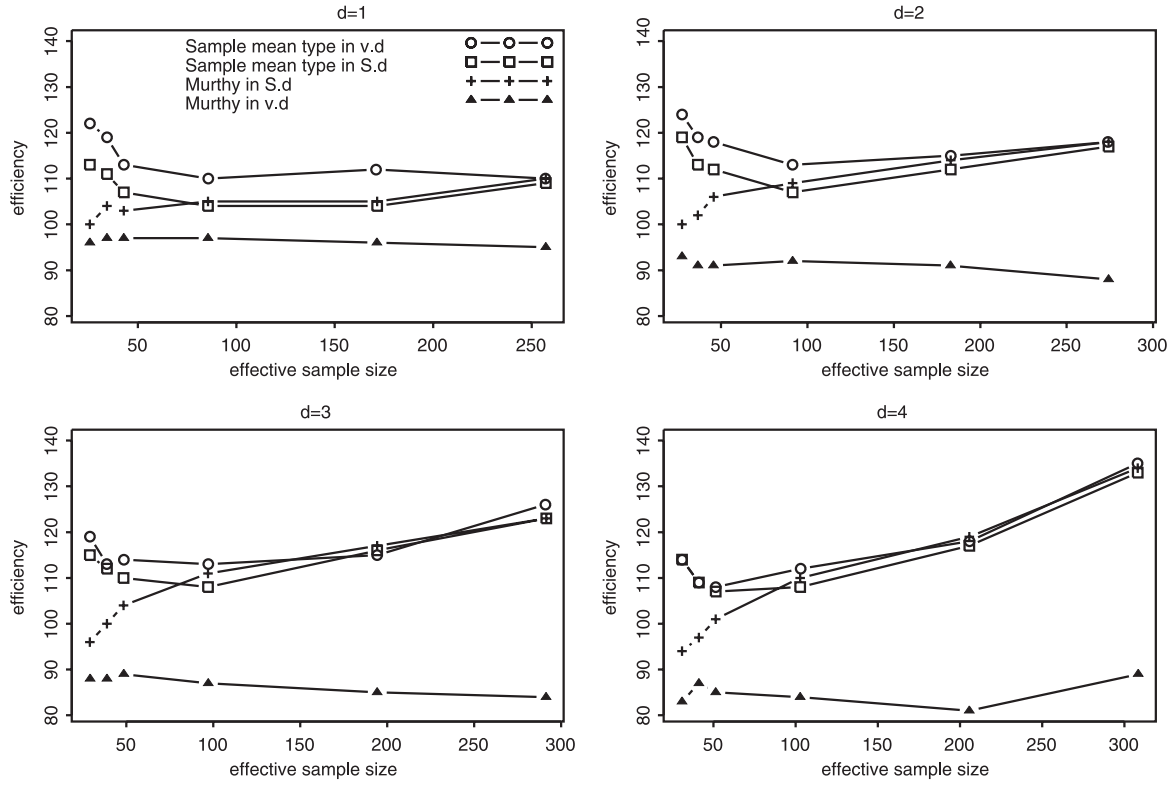

Figure 6. Efficiency of the sample mean type and Murthy's estimators in the variable sample size designs in the SZM population. The first phase sample size in each stratum is represented respectively as the set {3, 4, 5, 10, 30}.

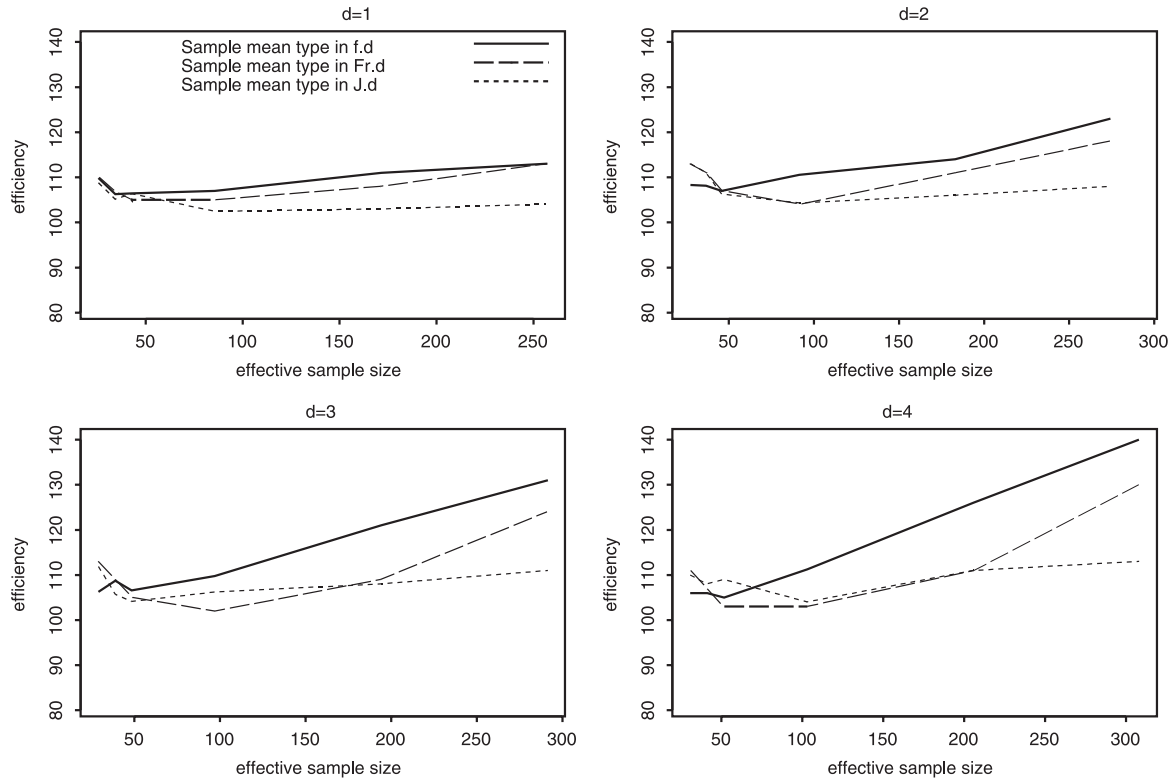

Figure 7. Efficiency of the sample mean type estimator in the fixed sample size designs in the SZM population. The first phase sample size in each stratum is represented respectively as the set {3, 4, 5, 10, 20, 30}.

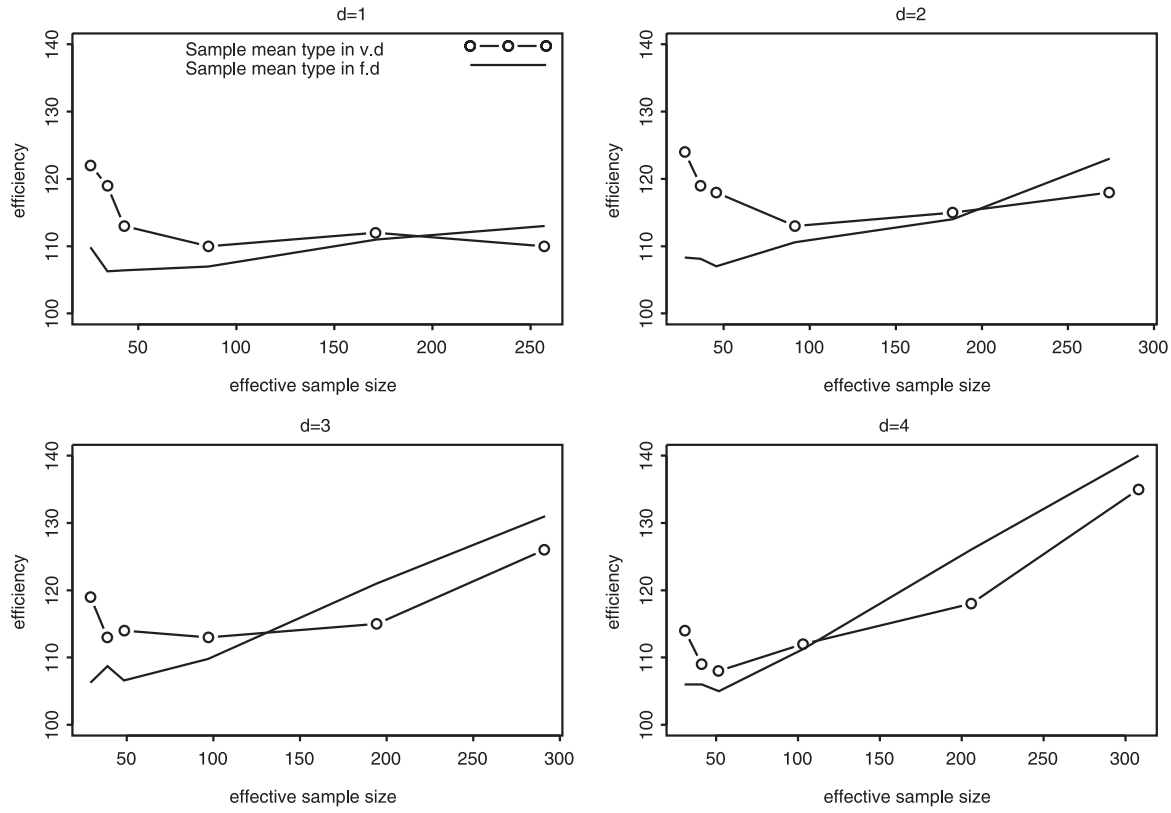

Figure 8. Efficiency of the best estimators in all designs in the SZM population. The first phase sample size in each stratum is represented respectively as the set {3, 4, 5, 10, 20, 30}.

identical efficiencies. They are also more efficient than estimator of Jolly–Hampton's design. In Figure 7, the sample mean type estimator for the fixed sample size design is more efficient than others.

For both the fixed and variable sample size designs, since the sample mean type estimator for the fixed and variable sample sizes were more efficient than estimators of introduced designs, we plot them in Figures 5 and 8. The variable sample size design was generally more efficient than the fixed sample size design at smaller effective sample sizes and the fixed sample size design was more efficient at larger effective sample sizes. Comparing the variable and fixed sample size designs, the results displayed in tables suggest that up until the first phase the allocation is about $1/4$ of the size of the population, and the variable size design is the more efficient of the two. When the first phase allocation is approximately more than $1/4$ of the size of the population, the fixed size design is more efficient.

It turns out that Murthy's estimator ($\hat{\tau}_{\mathrm{M}}$), which is derived from Rao-Blackwell procedure, is not an efficient estimator in our study. Results of Brown *et al.* [6] showed that this estimator can be efficient for very rare population. However, the weight $\hat{p}_h$ in $\hat{\tau}_{\mathrm{M}}$ depends on the first phase sample only, which can be a justification for not being an efficient sampling, especially when the first phase sample size is a small proportion of the final sample size.

In Table 3, for each estimator the relative bias in the fixed sample size design is smaller than the relative bias of that estimator in the variable sample size design. Relative bias will increase as multiplier d increases or the first phase sample size n_{h1} in each stratum is small, for all estimators. For $d < 5$ and $n_{h1} > 2$ the bias was negligible.

Acknowledgements

We thank Dr K.H. Low for providing us the moon data and its figure. Dr Salehi and Moradi's works were partially supported by the CEAMA of Isfahan University of Technology.

References

[1] W.G. Cochran, *Sampling Techniques*, 3rd ed., Wiley, New York, 1977.
[2] S.K. Thompson and G.A.F. Seber, *Adaptive sampling*, Wiley, New York, 1996.
[3] R.I.C.C. Francis, *An adaptive strategy for stratified random trawl survey*, N. Z. J. Mar. Freshwater Res. 18 (1984), pp. 59–71.
[4] G.M. Jolly and I. Hampton, *A stratified random transect design for acoustic surveys of fish stochs*, Can. J. Fish. Aquat. Sci. 47 (1990), pp. 1282–1291.
[5] M.M. Salehi and D.R. Smith, *Two-stage sequential sampling*, J. Agric. Biol. Environ. Stat. 10(1) (2005), pp. 84–103.
[6] J.A. Brown, M.M. Salehi, M. Moradi, G. Bell, and D. Smith, *Adaptive two-stage sequential sampling*, Popul. Ecol. 50 (2008), pp. 239–245.
[7] D.G. Horvitz and D.J. Thompson, *A generalization of sampling without replacement a finite universe*, J. Am. Statist. Assoc. 47 (1952), pp. 663–685.
[8] M.C. Christman, *Adaptive two-stage one-per-stratum sampling*, Environ. Ecol. Stat. 10 (2003), pp. 43–60.
[9] M. Moradi, M.M. Salehi, and P.S. Levy, *Using general inverse sampling design to avoid undefined estimator*, J. Probab. Statist. Sci. 5(2) (2007), pp. 137–150.
[10] K.H. Low, G. Gordon, J. Dolan, and P. Khosla, *Adaptive sampling for multi-robot wide-area exploration*, in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA'07), 2007, pp. 755–760.
[11] C.E. Särndal, B. Swensson, and J. Wretman, *Model Assisted Survey Sampling*, Springer Verlag, New York, 1992.

Appendix

In order to show that the bias of $\hat{\tau}_s$ is negative, we prove the expectation of the ratio of units satisfying the condition C in the sample is smaller than the ratio in the population. The ratio in the population and sample are, respectively, k/N and $(l_1 + l_2)/(n_1 + dl_1)$, the subscript h is eliminated for simplicity.

We first show that the function $f(l_1) = (l_1 + E_2(l_2))/(n_1 + dl_1)$ is a convex function where $E_2(.)$ is the expectation, given the first phase sample.

$$f(l_1) = \frac{l_1 + E_2(l_2)}{n_1 + dl_1} = \frac{l_1 + dl_1K - l_1/(N - n_1)}{n_1 + dl_1} = \frac{1}{N - n_1}\frac{(N - n_1 + dK)l_1 - dl_1^2}{n_1 + dl_1}$$

$$\begin{aligned}\triangle(f(l_1)) &= f(l_1 + 1) - f(l_1) = \frac{1}{N - n_1}\frac{(N - n_1 + dK)(l_1 + 1) - d(l_1 + 1)^2}{n_1 + d(l_1 + 1)} \\ &\quad - \frac{1}{N - n_1}\frac{(N - n_1 + dK)l_1 - dl_1^2}{n_1 + dl_1} \\ &= \frac{1}{N - n_1}\frac{-d^2l_1^2 - (2dn - d^2)l_1 - dn + (N - n_1 + dK)n}{(n_1 + dl_1)(n_1 + d(l_1 + 1))}\end{aligned}$$

$$\begin{aligned}\triangle^2(f(l_1)) &= \triangle(l_1 + 1) - \triangle(l_1) \\ &= \frac{1}{N - n_1}\frac{-d^2(l_1 + 1)^2 - (2dn - d^2)(l_1 + 1) - dn + (N - n_1 + dK)n}{(n_1 + d(l_1 + 1))(n_1 + d(l_1 + 2))} \\ &\quad - \frac{1}{N - n_1}\frac{-d^2l_1^2 - (2dn - d^2)l_1 - dn + (N - n_1 + dK)n}{(n_1 + dl_1)(n_1 + d(l_1 + 1))} \\ &= \frac{1}{N - n_1}\frac{-2dn(N + dK)}{(n_1 + dl_1)(n_1 + d(l_1 + 1))(n_1 + d(l_1 + 2))} < 0\end{aligned}$$

Using the Jenssen inequality, we have

$$\begin{aligned}E\left(\frac{l_1 + l_2}{n_1 + dl_1}\right) &= E_1\left(E_2\left(\frac{l_1 + l_2}{n_1 + dl_1}|l_1\right)\right) = E_1\left(\frac{l_1 + E_2(l_2)}{n_1 + dl_1}\right) \leq \frac{E(l_1 + l_2)}{E(n_1 + dl_1)} \\ &= \frac{n_1K/N(1 + d(K - n_1K/N)/(N - n_1))}{n_1 + dn_1K/N} = \frac{K}{N}\left(\frac{1 + dK/N}{1 + dK/N}\right) = \frac{K}{N}.\end{aligned}$$